A nice surprise? Predictive processing and the active pursuit of novelty

Phenom Cogn Sci
DOI 10.1007/s11097-017-9525-z

A nice surprise? Predictive processing and the active pursuit of novelty

Andy Clark[1,2]

Abstract Recent work in cognitive and computational neuroscience depicts human brains as devices that minimize prediction error signals: signals that encode the difference between actual and expected sensory stimulations. This raises a series of puzzles whose common theme concerns a potential misfit between this bedrock informationtheoretic vision and familiar facts about the attractions of the unexpected. We humans often seem to actively seek out surprising events, deliberately harvesting novel and exciting streams of sensory stimulation. Conversely, we often experience some wellexpected sensations as unpleasant and to-be-avoided. In this paper, I explore several core and variant forms of this puzzle, using them to display multiple interacting elements that together deliver a satisfying solution. That solution requires us to go beyond the discussion of simple information-theoretic imperatives (such as 'minimize long-term prediction error') and to recognize the essential role of species-specific prestructuring, epistemic foraging, and cultural practices in shaping the restless, curious, novelty-seeking human mind.

Keywords Prediction · Prediction error · Surprise · Novelty

This paper was drafted during a period of sabbatical leave (Autumn 2016) kindly granted by the University of Edinburgh, and completed as part of ERC Advanced Grant XSPECT - DLV-692739. Thanks to two anonymous referees for helpful comments on an earlier version.

✉ Andy Clark
andy.clark@ed.ac.uk

[1] School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, Scotland, UK

[2] Department of Philosophy, Macquarie University, Sydney, Australia

Published online: 27 September 2017

1 Predictive processing

Predictive processing[1] (henceforth, PP) depicts perception, cognition, and action as the closely woven products of a single kind of inferential process. That process ('active inference' in the sense of Friston et al. 2010) aims at the progressive reduction of organism-salient prediction error signals. The computational basis of perception, cognition, and action (if this ambitious story is on track) involves only three 'basic elements' – predictions (flowing from a long-term multi-level 'generative model'), prediction error signals (calculated relative to active predictions), and the estimated, context-varying 'precision' of those prediction error signals. Precision estimates the inverse variance of a prediction error signal – in other words, it sets error bars around an error signal according to its currently estimated importance or reliability. High-precision errors enjoy greater post-synaptic gain and (hence) increased influence. Conversely, even a large prediction error signal, if it is assigned extremely low precision, may be rendered systemically impotent, unable to drive learning or further processing.

The human brain, PP here suggests, commands a rich, integrated model of the worldly sources of sensory inputs, and uses that long-term model to generate on-the-spot predictions about the probable shape and character of current inputs. The rich, integrated (generative) model takes a highly distributed form, spread across multiple neural areas that may communicate in complex context-varying manners. The model itself is installed by some combination of evolutionary bequest and ongoing lifetime learning. But this adds nothing further to the onboard equipment since lifetime learning is itself enabled by the ongoing attempt at prediction. To predict the next word in a sentence, you would be helped by a knowledge of grammar. But one way to learn a surprising amount of grammar, as work in machine learning clearly demonstrates, is to try repeatedly to predict the next word in a sentence, adjusting your future responses in the light of past patterns. The brain thus uses the prediction task to bootstrap its way to the structured world-knowledge that is later used to generate better (and better) predictions.[2]

Prediction-driven learning, when implemented using hierarchical (hence multi-level) machinery automatically uncovers structure at multiple scales of space and time. Each higher level attempts to predict the sensory signal as it appears at the level below, and does so by learning about patterns in the lower level responses. Higher levels, in a language processing task, might thus specialize in predictions that involve word-level knowledge, while lower ones would specialize in predictions that involve letter-level knowledge, still lower ones in predictions that turn on stroke-level knowledge, and so

[1] Predictive Processing (Clark 2013a, b, c) is also known as hierarchical predictive coding and as 'active inference'. The former label stresses the layered, roughly hierarchical, organization of these systems and their use of the data compression technique known as 'predictive coding' while the latter is most closely associated with their more recent extensions into the domain of action and motor control. For discussion of these labels and of the (weakly) hierarchical assumption, see Clark (2016).

[2] The use of multi-layer neural nets, engaged in a prediction task, to acquire approximations to grammatical knowledge in simplified domains is well established (Elman 1991, 1993; Frank (2006)). This shows that the prediction task can indeed be used to bootstrap the learning of regularities in language. This falls short (as an anonymous referee correctly notes) of showing that the full complexities of a natural grammar could be learnt in this way. But this is not a problem, since predictive processing is silent on the important question of how much innate bias may be involved in language learning. For discussion, see Clark (2013c).

on. Higher level guesses are then used to contextualize and inform lower-level responses, while lower level responses inform the higher-level guessing. This give and take enables higher-level guesses at what the word is to resolve ambiguities at the letter level, for example between a poorly written 'ev' and a 'w' (for this example, see Friston 2002 and discussion in Clark (2016) chapter 5)). PP thus delivers a multi-scale grip on the worldly sources of structure in the sensory signal. This means (Kiebel et al. 2009) that higher levels come to specialize in predicting events and states of affairs that are 'built up' from the kinds of features and properties predicted by lower levels. The message-passing schema by which this is accomplished is one in which predictions flow sideways (within a level) and downwards, so that each level is attempting to use what it knows to predict the evolving pattern of activity at the level below. Notice that activity at each level is thus informed by the predictive attempts of the level immediately above. Each level effectively treats activity at the level below as if it were raw sensory evidence, and seeks to predict that evidence as it arrives. This will be important later in our treatment.

Prediction error signals then encode whatever is currently not accounted for by the top-down (and sideways) flow of prediction. As previously mentioned, to make the best and most flexible use of that flow of prediction error PP architectures simultaneously estimate the precision of the prediction error signal itself. This enables different circumstances to render different prediction error signals important, and may mandate different balances between processing in different brain regions and between top-down prediction and incoming sensory evidence. Reading a puzzling passage in an instruction manual while listening to music, I may need to devote my attention to the visually presented words, temporarily ignoring or down-regulating the impact of information from the auditory stream. In PP terms, that means increasing the estimated precision (hence the impact) of prediction errors regarding the manual-reading task. Trying to hear your words in a noisy room requires me to rely heavily upon my top-down predictions to fill in evidential gaps. But looking at a rare coin under a bright light implies taking small variations in the sensory signal very seriously indeed, the better to spot a well-made fake. In all such cases, the variable precision-weighting of prediction error provides a flexible means to adjust the balances of power according to task and context. Implemented by multiple means in the brain,[3] flexible precision-weighting thus renders these architectures astonishingly fluid and context-responsive (Clark 2016).

Finally, action itself (if PP is correct) is accomplished by a canny use of the same resources. The core idea (Friston et al. 2010) is that there are two ways for brains to match their predictions to the world. Either find the prediction that best accounts for the current sensory signal (perception) or alter the sensory signal to fit the predictions (action). If I predict I am seeing my cat, and error ensues, I might recruit a different prediction (e.g. 'I am seeing the computer screen'). Or I might move my head and eyes so as to bring the cat (who as it happens is right here beside me on the desk) into view. Importantly, that flow of action can *itself* be brought about, PP suggests, by a select sub-set of predictions – prediction of the (trajectory of) proprioceptive consequences that would ensue were the desired action to be performed. This turns out to be a computationally efficient (Friston 2011b) way of implementing motor commands.

[3] For example, slow dopaminergic modulation, and faster time-locked synchronies between neuronal populations. For discussion, see (Engel et al. 2001; Feldman and Friston 2010).

Action here results from a kind of self-fulfilling prophecy in which predicting a select subset of the sensory consequences of an action serves to bring the action about. Notice, however, that action also provides a potent means of testing the current hypothesis, for example disambiguating between equiprobable hypotheses (Seth (2015), Friston et al. (2012a)). The resulting picture is one in which perception and action are complementary manifestations of a single adaptive regime, whose core operating principle is the reduction of precise, organism-salient prediction error.

2 Three versions of the 'darkened room' worry

The most prominent concerns about the PP framework are all variants on questions concerning scope. Can this slim prediction-driven tool-kit really assemble and account for all the richness and variety of a human cognitive life? A popular way to press this worry is via the so-called 'Darkened Room' scenario (Friston et al. 2012b). Darkened Room scenarios come in various flavours. Some are designed to highlight the apparent *unattractiveness* of situations in which prediction error might seem to be minimized, while others depict the apparent *attractiveness* of situations in which prediction error might seem to be increased.

Version 1 can be labeled "The Death Trap". The worry here is that the cognitive imperative of prediction error minimization can seem like a recipe for adaptive disaster. If brains like ours are driven only to minimize prediction error signals, why don't creatures like us simply find a nice dark corner (providing fully predictable, meager, unvarying patterns of sensory stimulation) and stay there, slowly growing weaker and then dying? Such creatures would surely display what Willard Van Orman Quine nicely described as a 'pathetic but praiseworthy tendency to die before reproducing their kind' [4] – In this case, by being the victims of their own prediction error minimizing success.

Version 2 can be called "The Boredom Trap". The idea here is that a common response to the Death Trap worry – one which we will briefly rehearse below - leads to an equally pernicious variant. On this variant, the prediction error minimizing routine has the resources to ensure that we stay alive, but dooms us to sample the world so as to actively harvest the sensory streams we expect.

Version 3 is the hardest to describe, but (I hope to show) the most potentially damaging. Let's call it "The Merely Modest Exploration Trap". It arises insofar as plausible responses to the first two variants involve appeals to the adaptive value of a certain amount of exploration – for example, it's value in reducing long-term uncertainty and insuring against volatile environments. But such appeals still seem to fall short of providing a full or distinctive account of the deep positive attractions of novelty, play, and exploration. Yet we humans (so the worry goes) appear to value novelty and exploration in ways not obviously reducible to instrumental value. Indeed, whole traditions in art and entertainment have been built around the perceived value of experiential surprise. The Darkened Room worries thus bring us face to face with one of the major questions confronting any account of core sub-personal mechanisms – the question of its 'fit' with conscious, personal-level experiential facts.

[4] The phrase occurs as part of a discussion of inductive reasoning, in the essay "Natural Kinds" in the collection *Ontological Relativity and Other Essays* (1969), p. 126.

In sum, the worry – here appearing in three distinct yet closely related guises - is that the core imperative to reduce long-term prediction error looks inconsistent with much that we humans deem valuable and attractive, rendering the PP story at best radically incomplete.[5]

3 Avoiding the simple death trap

The Death Trap worry has a standard, but I think importantly underspecified, solution. The solution (see, for example, (Friston 2011a) is to claim that creatures like us simply do not 'expect' to sit still and die in dark nutrition-less corners. Even well-adapted darkness-dwellers (troglodytes) would 'predict' motion, foraging, and feeding, and those predictions would (courtesy of the standard PP apparatus of 'active inference' described earlier) help bring such motions and foraging expeditions about. In much the same vein, it should be noted (Hohwy 2013) that there is no inconsistency between active exploration (e.g. searching for food in new places) and the minimization of prediction error, as long as we are careful about the various time-scales over which these occur. If we 'expect' sufficient food in some deep, bedrock way, then we may reduce long-term surprise relative to that expectation by taking steps to ensure that food sources remain available even in changing and volatile environments.[6]

Another way to express this is by distinguishing 'surprise', in its usual experiential sense, from 'surprisal' – the information-theoretic construct representing the difference between sensory evidence and predictions made on the basis of a model of the world.[7] Thus we read that:

> Put simply, (most) agents would find it highly surprising to be incarcerated in a dark room and would thus generally try to avoid that state of affairs. More formally, there is a fundamental difference between the intuitive meaning of "surprise" in terms of unpredictable sensory input and surprise (in information theoretic terms) under a particular model of the world. Finding ourselves in a dark room (and being subject to a surprising sense of starvation and sensory

[5] A tempting response at this point would be simply to insist that PP offers a staunchly sub-personal story that is therefore fully consistent with any personal-level account. Such a response might be bolstered by reflection on the important distinction (see section 3 following) between personal-level (conscious, experiential) surprise and sub-personal 'surprisal'. I choose not to take this direct route, because it seems inconceivable (from my broadly naturalistic perspective) that personal-level facts should float free of the swathes of sub-personal activity upon which they supervene. My view (which I will not attempt further to defend here) is thus that each discourse constrains the other, such that apparent conflicts must eventually be resolved so as to reveal relations, however complex, between our conscious experiences and the complex weave of probabilistic sub-personal prediction. All that is, of course, consistent with the view that personal-level descriptions serve special purposes, and may carve nature in ways alien to other discourses. What is required is simply that our personal-level profiles should be intelligible in the light of good mechanistic models of their physical underpinnings. It is apparent threats to this intelligibility that the current paper aims to dissolve. For some useful discussion, see Drayson (2014).

[6] For explorations of many of these issues, cast in terms of the more familiar 'exploit-explore' balances required for adaptive success, see Hammerstein and Stevens (Eds.) (Hammerstein and Stevens 2012). For a rich PP-style exploration of the issues, see Friston et al. (2016).

[7] See (Tribus 1961)

> deprivation) is a highly surprising state, even though it represents an environment with maximally predictable sensory input. (Schwartenbeck et al. 2013) p. 2

There is something right about all these moves. Quite generally, animals like us live in a changing and challenging world and so must adopt quite complex strategies (Friston 2010) to stay within their species-specific windows of viability. In this way, Friston suggests, our 'neural expectations' may come to include expectations of 'itinerant trajectories' mandating change, exploration, and search. We 'expect' to sometimes engage in random environmental search as a means of entering into adaptively valuable states. To put it crudely, we randomly sample because - qua evolved organisms - we 'expect' to discover food, mates, or water at some point during the expedition.

But such responses open the door to a different kind of worry. It is that the notions of prediction and expectation may now be being stretched beyond their proper limits. Do we really predict or expect constant supplies of food and water, and are we really 'surprised' (even in a technical, information-theoretic sense) when, halfway through a local famine, such supplies are no longer to be found? . It is this very broad sense of 'prediction' that (Anderson and Chemero 2013) suggest threatens to trivialize PP's claim to offer a grand unifying theory of the brain. To this, it may be responded that all that is meant is that the animal has priors that specify the various forms of escape. But the issue then becomes how much explanatory work such appeals to priors can do – for any behaviour whatsoever can be cast as optimal with respect to some set of (perhaps bizarre) priors, as formally demonstrated by Brown (1981).

Can we do better? A clue emerges in work on 'interoceptive predictive coding' (see e.g. (Seth 2013; Barrett and Simmons 2015; Pezzulo 2014). This work highlights an under-appreciated feature of the total sensory stream that the agent is trying to predict. That feature is the stream of information specifying (via dense vascular feedback) the physiological state of the body. That includes the state of the gut and viscera, blood sugar levels, temperature, and a great deal more- see (Craig 2014) for a full review. What happens when a unified multi-level prediction engine crunches all that intero-ceptive information together with exteroceptive information specifying states of affairs in the world? Such an agent has a predictive grip on multi-scale structure in the external world. But that multi-layered grip is now superimposed upon (indeed, co-computed with) *another* multi-layered predictive grip – a grip on the changing physiological state of her own body. And these clearly interact. As your bodily states alter, the *salience* of various worldly opportunities alters too. Such estimations of salience are written deep into the heart of the predictive processing model, where they appear (as we just saw) as alterations to the weighting (the 'precision') of specific prediction error signals. As those estimations alter, you will act differently, harvesting different streams of extero-ceptive and interoceptive information, that in turn determine subsequent actions, choices, and bodily states.

Add to that picture recent work (Kanai et al. 2015) that foregrounds the role of sub-cortical influences in the assignments of precision that determine perception and action. Here sub-cortical structures (especially the thalamus, and within it, the pulvinar) seem poised to play a special and crucial role. Understanding that role requires us to move beyond what ((Pessoa 2014) p.11) describes as the 'corticocentric' image in which evolutionary older sub-cortical structures are dominated and controlled by the more

recent cortical overlay. Instead, we are led to endorse an 'embedded' view (Pessoa (op cit)) according to which cortical and sub-cortical states and activities change in a coordinated fashion characterized by ongoing patterns of mutual influence. Thus the pulvinar, to take the most studied example, is now thought to play a key role in gating (by the potent tool of variable precision-weighting (Kanai et al. 2015)) core aspects of cortico-cortico communication.

The overall effect should be to seamlessly merge sub-cortical and cortical influences during action and action-selection. Sub-cortical influences here bias (to use the term favoured by Pessoa) large-scale neural patterns towards signals that are *biologically valuable* – those accorded high precision within the PP scheme. Consistent with this, (Barrett and Bar 2009) provide a physiologically detailed account of how affect, action, perception and prediction are also constantly co-computed, courtesy of ongoing looping exchanges with 'core-affect' areas such as the OFC, ACC, insula and amygdala. All this keep us from straying too far from adaptive 'set points' reflecting basic organismic needs for food, exercise, company, and water. Talk of 'expecting to find' such staples is now revealed as just a kind of innocent summary shorthand for complex distributed webs of prediction and (especially) precision- estimation. This delivers a much more substantive unpacking of the idea of an 'embodied predictive model', and one relative to which the Death Trap gets no purchase.

4 Beyond boredom

There remains, however, a subtly different version of the Death Trap worry – the version dubbed 'the Boredom Trap'. It is nicely presented in the following passage:

> If our main objective is to minimize surprise over the states and outcomes we encounter, how can this explain complex human behavior such as novelty seeking, exploration, and, furthermore, higher level aspirations such as art, music, poetry, or humor? Should we not, in accordance with the principle, prefer living in a highly predictable and un-stimulating environment where we could minimize our long-term surprise? (Schwartenbeck et al. 2013, p. 2)

Schwartenbeck et al. have a positive story to tell here, and I return to it in section 7 below. Similar worries – though without the accompanying positive story - have been raised by (Froese and Ikegami 2013). These authors suggest that the long-term minimization of prediction-error might mandate 'stereotypic self-stimulation, catatonic withdrawal from the world, and autistic withdrawal from others'.

A first – but ultimately insufficient – response is to notice that many of the remedies outlined in 3.2 still apply. The experiential attractiveness of just-right doses of novelty and (agentive) surprise may thus be implemented by complex cortico-sub-cortical loops that implicate multiple systems of reward and value. These systems of reward and value are, we saw, neatly accommodated within the sub-cortically expanded PP scheme, where high-precision prediction errors track valuable (hence preferentially action-entraining) states of affairs. It is also worth stressing the potential species-level advantages of a spectrum of individually differing tendencies towards play and exploration,

as realized by these complex webs of cortico-sub-cortical influence. This too plausibly represents a good evolutionary strategy for optimally combining, at the population level, the exploitation of known resources with the search for new ones. In other words, having a spread of individual differences in tendencies to exploit and explore presumably makes for a more robust population – one in in which some, but not all, individuals will be more inclined to take greater risks, and to forage further afield.

Simple versions of the Boredom Trap can thus be avoided, individual differences accommodated, and the fundamental attractions of play and novelty preserved. This is progress. It plugs varying tendencies towards play and exploration right into the heart of the predictive economy, allowing some prediction error minimizing agents deliberately to seek out adaptively valuable doses of complexity and challenge. It does so, moreover, without trivializing the notion of prediction itself. But it is not enough.

5 What information-theoretic imperatives can't do

Even this, however, may seem unsatisfying to the determined critic who now reformulates her worry by suggesting that these prediction-error-minimizing agents exhibit at most a modest and instrumentally-motivated tendency towards play, exploration, and the search for novel experiences. These prediction error minimizing agents remain locked, it seems, into an information-theoretic journey whose guiding principle is in some way unacceptably conservative. It is a journey which, if successful, will be marked only by the attainment of *expected goals and meta-goals* (including, for advanced agents, expected doses of exploration, epistemic foraging, and novelty-seeking). But such agents, it may be feared, remain fundamentally cognitively conservative. Doesn't this leave unexplained the basic human need for 'self-actualization and personal growth' (Maslow 1943))? Such commonly felt needs are better captured, it may be suggested, by considering open-ended objectives such as increasing the richness of the events of which we can make sense, rather than by more conservative information-theoretic ideas such as those that appear centered upon survival and homeostasis.

This, then, is the third, final, and most genuinely challenging incarnation of the Darkened Room worry – the one we dubbed the 'Merely Modest Exploration Trap'. Prediction error minimizing agents are driven – or so the worry goes - by a fundamental information-theoretic goal that is itself inimical to human flourishing.[8] For such agents, the ultimate information-theoretic goal is a state in which there is zero prediction error. This looks diametrically opposed to oft-lauded goals such as continued personal growth and self-actualization (see Maslow 1943; Seligman et al. 2013). How, if at all, are we to reconcile such expansive visions of human flourishing with the information-theoretic goal of prediction error minimization?[9]

One response (which we should reject, for reasons that will emerge shortly) is to seek some subtly different core information-theoretic imperative. In just this vein,

[8] Thanks to Bill Phillips for impressing this worry upon me.

[9] I suspect – and thanks to a helpful referee for bringing this out – that the sceptic can always find some notion of 'flourishing' that any specific positive story fails to deliver. This is because there are, as we are seeing, multiple tricks and ploys that conspire to help us generally avoid the dark room scenarios The take home message is that there are many traps, and just as many escape routes.

(Little and Sommer 2013) suggest that we should shift our attention from the minimization of prediction error to the maximization of mutual information (hence the maximisation of prediction success). Such an agent would seek to maximize the mutual information (see also (Phillips 2013)) between an internal model of estimated causes and the sensory inputs. Minimizing entropy (prediction error) and maximizing mutual information (hence prediction success), Little and Sommer argue, will each reduce prediction error but would differ in how they select actions. A system that seeks to maximize mutual information won't, they argue, fall into the dark room trap. For it is driven instead towards some sweet spot between predictability and complexity and will "seek out conditions in which its sensory inputs vary in a complex, but still predictable, fashion".

Compelling empirical evidence for such a profile includes work by (Kidd et al. 2012) who conducted a series of experiments with 7- and 8-month-old infants measuring attention to sequences of events of varying (and well-controlled) complexity. Infant attention, they found, was characterized by what they dub a 'Goldilocks Effect', focusing upon events presenting an intermediate degree of predictability—neither too easily predictable, nor too hard to predict. The probability of an infant looking away was thus greatest when complexity (calculated as negative log probability) was either very high or very low. The functional upshot, Kidd et al. suggest, is that 'infants implicitly seek to maintain intermediate rates of information absorption and avoid wasting cognitive resources on overly simple or overly complex events' (Kidd et al. 2012, p. 1). Such tendencies to seek out 'just-novel-enough' situations are a good candidate for some form of innate specification, since they would cause active agents to self-structure the flow of information in ways ideally suited to the incremental acquisition and tuning of a rich and informative generative model of their environment. More generally still, agents that inhabit complex, changing worlds would be well-served by a variety of policies that drive them to explore those worlds, even when no immediate gains or rewards are visible. This kind of 'epistemic foraging' is itself mandated by the pressure to reduce prediction error across longer and longer stretched of time.

Related proposals have been made by (Oudeyer and Smith 2016) who suggest that the core information-theoretic principle should be one that drives organisms to experience ongoing improvements, manifesting as a continual decrease in prediction errors, rather than seeking simply to minimize prediction errors. This would likewise rule out Darkened Room scenarios in which there is no trajectory of improvement of prediction. Oudeyer and Smith test their hypothesis in a simple robotic experiment, in which the drive to experience an ongoing reduction in prediction error implements a version of 'curiosity-drive learning'. The robots actively selected experiences that delivered progressive decreases of prediction error, yielding a kind of 'self-organized epigenesis'. In these experiments, Oudeyer and Smith argue, "Progress in learning in and for itself generates intrinsic rewards and an action selection system directly aims to maximize this reward" (op cit p.2).

Curiosity-driven learning and exploration are, I agree, among the key features that characterize a great deal of animal behaviour. They are features moreover, that seem especially well-developed in the human case, leading Oudeyer and Smith to speculate that "the dominance of curiosity in the motivational hierarchy may be key to the emergence of wide-ranging domain-specific knowledge in humans as compared to other species" (op cit p.9). But curiosity-driven learning and exploration are not best

seen as *alternatives* to the core information-theoretic imperative of reducing long-term prediction-error. Instead – or so I suggest - they reflect the operation of that very principle, as it plays out within species-specific webs of neuro-anatomical structure and – in the human case especially - cultural constraint.

To see this, it helps first to appreciate that no information-theoretic target can, in principle, be immune from what I shall dub 'information-theoretic subversion'. Thus take a system aiming to maximize prediction success. Such a system will, to be sure, exhibit a certain kind of ongoing expansion. For to maximize prediction success, it will continually extend its reach to encompass as new states and patterns. But there will be many ways to do this, most of which would not represent any kind of human flourishing or self-actualization. Such an agent might, for example, continuously re-arrange a finite (but lifetime inexhaustible) stock of small coloured tiles into larger and larger patterns, inducing then learning new predictive relationships (if I look at this pattern then move my eyes left, this is what I will see) at every stage. Mutual information between the agent's neural economy and the environment is now increasing at every step. But this, surely, is just another way of becoming sucked into a (somewhat more psychedelic) version of the darkened-room. Such an agent, we may suppose, is constrained by more ancient homeostatic drives to maintain her integrity, to eat, drink, and mate. But her quality of life, her degree of flourishing, looks every bit as limited as that of the prediction-error-minimizing agent. A similar scenario can be constructed for the agent driven to experience an ongoing reduction in prediction error. Such an agent could be 'hijacked' by a darkened room containing only a computer monitor and an endless supply of simple (but not too simple) puzzles, each of which allows a steady trajectory of improvement until the next one pops up on the screen.

The point here is perfectly general (Clark 2013b)). Any purely information-theoretically specifiable goal will be subvertable, it seems, by the right set of environmental conditions. For any such agent, no matter *what* information-theoretic quantity or quantities they are aiming to maximize or minimize, there will be a scenario, consistent with that imperative, that looks inconsistent with true flourishing. Some imaginable scenario, tailored to whatever information-theoretic recipe is on offer, will provide an endless but in some intuitive sense boring or trivial supply of whatever quantity is being maximized or minimized – and this will be true no matter how complex or systematically changing the information-theoretic goal, and whether or not that goal includes medicinal doses of 'required randomness'. Any information-theoretically specifiable target, I am suggesting, will be subject to this kind of subversion.

6 Predicting in designer worlds

Such subversion does not, however, typically occur.[10] The reason for this, I suggest, is that we humans minimize prediction errors within the ecologically unique, self-engineered contexts of culture, technology, and linguaform exchange. This provides a second, and potent, means of avoiding the traps. It is our cultural practices themselves that conspire to render us humans so deeply exploratory and that more-or-less

[10] Near approximations include the tendency to addiction and various forms of obsessive behavior. For some nice discussion, see (Montague 2006; Friston 2012).

immunize us against information-theoretic subversion of even the 'merely modest' kind. Constrained and enabled to survive by our basic homeostatic economy, we humans – like other animals - induce complex multi-layered distributed world models. These models span cortical and sub-cortical systems, directly implicating action, play, and exploration in the ways sketched earlier. But as a kind of spin-off from all this deep (and language-enriched – see (Lupyan and Clark 2015)) model-building activity, we humans also began to construct complex social and physical environments. Our human minds are now marinated in the unique statistical baths of a succession of such designer worlds –worlds characterized by the complex traditions and practices of art, science, recreation, and literature. Those practices reliably spawn an open-ended set of new local goals and projects (think Pokemon Go). The skilled pianist has learnt to reduce prediction error with respect to complex melodies and motor repertoires, and the skilled mathematician with respect to properties and relations among numbers, theorems, and other constructs. But the musical and mathematical traditions within which they operate reflect the operation of cultural forces such as practices of writing, reflecting, disseminating, and peer review. Such practices – beautifully described by (Hutchins 2011) - repeatedly push us away from local equilibria, ensuring a steady diet of change, innovation, and challenge. Indeed, we may expect complex ratcheting effects here as altered environments install policies that drive the creation of further altered environments – a process nicely described by Heyes (2012) and further discussed in Clark (2016, Chapter 9). These powerful effects are further explored in work by 'enactivists' sympathetic to PP- for example, Rietveld and Kiverstein (2014), Bruineberg et al. (2016), Gallagher et al. (2013)). For a discussion of some issues that might (or might not) still separate these traditions, see Clark (2016) chapter 9, Madary (2015).

For present purposes, what matters is just that by designing and repeatedly redesigning our own environments, populating them with new books, paintings, theories, games, and practices, we humans continually move the goalposts for our own prediction-based learning. In this way cultural processes deliver new environments that set new targets for the same old prediction error minimizing machinery to cope with – much as if an animal were repeatedly to alter the niche with respect to which long-term prediction error needs to be reduced. An implication may be that true conceptual novelty, when it arises, is better explained as (at least in large part) a result of the framing and scaffolding of human activity by shifting cultural practices and changing sets of concrete constraints. For example, Wheeler (forthcoming) shows, using multiple real-world case studies, how 'external developments' such as new equipment and room acoustics can create physical spaces and opportunities that constrain individual exploration in ways that yield new forms of music and musical performance.[11] This also speaks to the observation[12] that cultural scaffolding (e.g. via immersive experience within a musical style or tradition) often seems to constrain novelty rather than (as I argue) enabling it. Both effects routinely occur, with conceptual novelty often resulting from changes to what might superficially seem like quite tangential factors – but ones that turn out to have deep impacts upon individual and collective innovation.

[11] Thanks to an anonymous referee for encouraging me to flag this implication.
[12] Thanks to an anonymous referee for pursuing this issue.

Finally, consider the suggestion (Schwartenbeck et al., op cit) that some evolved agents may acquire policies that positively value the opportunity to visit many new states. For such agents, the value of some current state is partially determined by the number of possible other states that it allows them to visit. But how do such policies get a foothold in the first place? The answer may again lie in the complex human-built environments of art, literature, and science. Predictive agents immersed in these kinds of designer environment learn to value (assign high precision to) states that make available wide ranges of new moves and outcomes, enabling scientists and designers to discover ideas that bring rewards that register in basic forms of neural circuitry (Montague 2006). Our evolving cultural practices thus piggyback upon ancient reward systems in ways that enable human populations to continuously seek out new ideas and perspectives.

Does all this threaten to revive the original puzzle, requiring there to be some information-theoretic 'goal' that our communal cultural explorations serve? I see no reason to think so. Instead, an operative information-theoretic imperative need only explain why creatures like us are built the way we are, with mechanisms that assign proximal value (here manifesting as precision) in ways that can serve any project that we have embraced. Against the enabling backdrop of the homeostatic machinery that keeps us within our windows of organismic viability, the shape and contents of the rest of our mental lives are determined by prediction-driven learning as it unfolds in the ecologically unique context of our many designer environments – mathematics, philosophy, fashion, dance, music, sailing, and mountain-climbing, to name but a few. Most of the heavy lifting, when it comes to explaining the shape and nature of the modern mind, is thus be done by our peculiar social and intellectual histories – histories replete (Clark 2013b) with chance and path-dependent unfolding.

7 Conclusions: Putting prediction error in its place

Darkened Room scenarios are among the most popular reasons for scepticism concerning the most extreme explanatory aspirations of predictive processing. Standard replies to such worries can seem ad hoc, asserting that evolved agents simply 'expect' to play, to explore, and to experience epistemically useful amounts of novelty and surprise. But things look better once when the PP story is expanded to include patterns of sub-cortical influence and complex training enivironments. Talk of 'expectations' of play and novelty then stand revealed as a kind of shorthand for these webs of interacting processes whose combined effect is to ensure that prediction-error minimizing agents keep busy, harvesting information and improving their overall model of embodied exchange with the world. In our remarkably social species, this process runs via designer environments such as art, literature, music, and science. Our prediction-error-minimizing brains are thus immersed in exotic statistical baths that enforce exploration and novelty-seeking in ways hitherto unknown among terrestrial animals.

References

Anderson, M. L., & Chemero, T. (2013). The problem with brain GUTs: Conflation of different senses of "prediction" threatens metaphysical disaster. *Behavioral and Brain Sciences, 36*(03), 204–205.

Barrett, L. F., & Bar, M. (2009). See it with feeling: Affective predictions during object perception. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 364*(1521), 1325–1334.

Barrett, L. F., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience, 16*(7), 1–11.

Brown, L. (1981). A complete class theorem for statistical problems with finite sample spaces. *The Annals of Statistics, 9*(6), 1289–1300.

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2016). The anticipating brain is not a scientist: The free-energy principle from an ecological-enactive perspective. *Synthese, 1-28*. doi:10.1007/s11229-016-1239-1

Clark, A. (2013a). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*, 181–204.

Clark, A. (2013b). Are we predictive engines? Perils, prospects, and the puzzle of the porous perceiver. *Behavioral and Brain Sciences, 36*(03), 233–253.

Clark, A. (2013c). Expecting the World: Perception, Prediction, and the Origins of Human Knowledge. *Journal of Philosophy, 110*(9), 469–496.

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press, NY.

Craig, A. D. (2014). *How do you feel? An interoceptive moment with your neurobiological self*. Princeton: Princeton University Press.

Drayson, Z. (2014). The personal/subpersonal distinction. *Philosophy Compass, 9*, 338–346.

Elman, J. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 7*, 195–224.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition, 48*, 71–99.

Engel, A. K., Fries, P., & Singer, W. (2001). Dynamic predictions: Oscillations and synchrony in top-down processing. *Nature Reviews Neuroscience, 2*(10), 704–716.

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience, 4*, 215. doi:10.3389/fnhum.2010.00215

Frank, S. L. (2006). Learn more by training less: Systematicity in sentence processing by recurrent networks. *Connection Science, 18*, 287–302.

Friston, K. (2002). Beyond phrenology: What can neuroimaging tell us about distributed circuitry? *Annual Review of Neuroscience, 25*, 221–250.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews. Neuroscience, 11*(2), 127–138.

Friston, K., 2011a. Embodied inference: or "I think therefore I am, if I am what I think ." In W. Tschacher & C. Bergomi, eds. *The implications of embodiment (Cognition and Communication)*. Imprint Academic, Exeter, UK pp. 89–125.

Friston, K. (2011b). Perspective what is optimal about motor control ? *Neuron, 72*(3), 488–498.

Friston, K. (2012). Policies and priors. In B. Gutkin & H. S. Ahmed (Eds.), *Computational neuroscience of drug addiction* (pp. 237–283). Springer New York: New York.

Friston, K. J., et al. (2010). Action and behavior: A free-energy formulation. *Biological Cybernetics, 102*(3), 227–260.

Friston, K., Adams, R. A., Perrinet, L., & Breakspear, M. (2012a). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology, 3*, 151. doi:10.3389/fpsyg.2012.00151.

Friston, K., Thornton, C., & Clark, A. (2012b). Free-energy minimization and the dark-room problem. *Frontiers in Psychology, 3*(May), 1–7.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., & Pezzulo, G. (2016). Active inference and learning. *Neuroscience and Biobehavioral Reviews*. doi:10.1016/j.neubiorev.2016.06.022.

Froese, T., & Ikegami, T. (2013). The brain is not an isolated "black box," nor is its goal to become one. *Behavioral and Brain Sciences, 36*(03), 213–214.

Gallagher, S., Hutto, D., Slaby, J., & Cole, J. (2013). The brain as part of an enactive system. *Behavioral and Brain Sciences, 36*(4), 421–422.

Hammerstein, P., & Stevens, J. R. (Eds.). (2012). *Evolution and the Mechanisms of Decision Making*. Cambridge: MIT Press.

Heyes, C. (2012). Grist and mills: On the cultural origins of cultural learning. *Philosophical Transactions of the Royal Society B, 367*, 2091–2096.

Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.

Hutchins, E., 2011. The role of cultural practices in the emergence of modern human intelligence. , (February 2008), *Philosophical Transactions of the Royal Society B 2008 363*, 2011–2019. doi:10.1098/rstb.2008.0003.

Kanai, R., et al. (2015). Cerebral hierarchies : predictive processing , precision and the pulvinar. *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences, 370*, 20140169.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS One, 7*(5), e36399.

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2009). Perception and hierarchical dynamics. *Frontiers in Neuroinformatics, 3*(July), 20.

Little, D. Y.-J., & Sommer, F. T. (2013). Maximal mutual information, not minimal entropy, for escaping the "Dark Room.". *Behavioral and Brain Sciences, 36*(03), 220–221.

Lupyan, G. & Clark, A., 2015. Words and the world: Predictive coding and the language-perception-cognition interface. *Current directions in Psychology ,24*(4), 279–284.

Madary, M (2015) Extending the Explanandum for Predictive Processing. In Open MIND. T. Metzinger and J. Windt, eds. Imprint. https://open-mind.net/papers.

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review, 50*(4), 370–396.

Montague, R., 2006. Why choose this book? : How we make decisions Penguin, New York.

Oudeyer, P. Y., & Smith, L. B. (2016). How evolution may work through curiosity-driven developmental process. *Topics in Cognitive Science*, 1–11.

Pessoa, L. (2014). Understanding brain networks and brain organization. *Physics of Life Reviews, 11*(3), 400–435.

Pezzulo, G. (2014). Why do you fear the bogeyman? An embodied predictive coding model of perceptual inference. *Cognitive, Affective, & Behavioral Neuroscience, 14*(3), 902–911.

Phillips, W. A. (2013). Neuronal inference must be local, selective, and coordinated. *Behavioral and Brain Sciences, 36*(03), 222–223.

Rietveld, E., & Kiverstein, J. (2014). A rich landscape of affordances. *Ecological Psychology, 26*(4), 325–352.

Schwartenbeck, P., et al. (2013). Exploration, novelty, surprise, and free energy minimization. *Frontiers in Psychology, 4*(October), 710.

Seligman, M. E. P., Railton, P., Baumeister, R. F., & Sripada, C. (2013). Navigating into the future or driven by the past. *Perspectives on Psychological Science, 8*(2), 119–141.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences, 17*(11), 565–573.

Seth, A.K (2015) The Cybernetic Bayesian Brain, in MIND collection. Open MIND. Metzinger, T. and Windt, J. eds. Imprint. https://open-mind.net/papers/the-cybernetic-bayesian-brain.

Tribus, M., 1961. Thermostatics and thermodynamics: An introduction to energy, information and states of matter, with engineering applications, Van Nostrand, NY.